2nd week of Jan: pres. of the project

I. The beginning: Monge's Problem

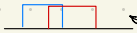
Take two sets of points with the same number of elements.

(MP) $\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} c(x_i, y_{\sigma(i)})$ (cost of transporting points of $\bar{x}$ to points of $\bar{y}$)

Spoiler: We found an algorithms in $O(n^3)$ - Hungarian, Auction (Bentsekas)

1D case:  $c(x,y) = h(x-y)$, h convex, e.g. $c(x,y) = |x-y|^p$, $p \geq 1$

- **Prop.**: For h convex, any optimal σ is monotonic (respects the sorting) $\Rightarrow$ algorithm in $O(n \log(n))$
Moreover, if h is strictly convex and points are disjoint, then $\exists!$ OT σ.

- **RK.**: Problem with $c(x,y) = |x-y|$  ← infinite solutions

- **Prop.**: $X = Y = \mathbb{R}^d$ (ambient space of points) $c(x,y) = \|x-y\|^p$
$p=1 \Rightarrow$ No crossing allowed

Lets generalise:

X a complete metric space (Polish space), separable. $\alpha \in \mathcal{M}(X)$, $\beta \in \mathcal{M}(Y)$ Borel measures = Radon measures (because Polish)

Notation for discrete measures: $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$, $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$

Push-forwards: $T: X \to Y$ some map, $T_\# : \mathcal{M}(X) \to \mathcal{M}(Y)$ linear

- **Def.**: $T_\#(\delta_x) = \delta_{T(x)}$, $T_\#\left(\sum_i a_i \delta_{x_i}\right) = \sum_i a_i \delta_{T(x_i)}$

- **Def. 1**: $T_\# \alpha(B) := \alpha(T^{-1}(B))$, $\forall B \subset Y$ measurable

- **Def. 2**: $\forall g \in \mathcal{C}_c(Y)$, $\int g(y) d\beta(y) := \int g(T(x)) d\alpha(x)$

- **Prop.**: if $T: X \xrightarrow{C^1} Y$ ($X, Y \subset \mathbb{R}^d$) is a diffeo. and α has density ρ_α w.r.t. λ then β has a density ρ_β w.r.t. λ and
$$\rho_\alpha(x) = |\det(\delta T(x))| \cdot \rho_\beta(T(x)).$$

Now:
$$(MP): \quad \alpha \in \mathcal{P}(X),\ \beta \in \mathcal{P}(Y) \qquad \inf_{T_\# \alpha = \beta} \int c(x, T(x)) d\alpha(x)$$

- **Rq.**: $T_\#$ can only diminish the size of the support.

· Prop. / RK.: if $\begin{cases} \alpha = \frac{1}{n}\sum_i a_i \delta_{x_i} \\ \beta = \frac{1}{n}\sum_i b_i \delta_{y_i} \end{cases}$, then $T_\# \alpha = \beta \Leftrightarrow T$ induces a permutation ($T(x_i) = y_{\sigma(i)}$)

· Theo.: (Brenier)
$X = Y = \mathbb{R}^d$, $c(x,y) = \|x-y\|_2^2$ (we can take manifolds and $d_{geod}(x,y)^p$, $p>1$), α has a density w.r.t Leb.
i) $\exists!$ OT map T^* (α-a.e.)
ii) T^* is the only $T = \nabla\psi$, ψ convex, $(\nabla\psi)_\# \alpha = \beta$.

· RK.: ψ convex, $T = \nabla\psi \Leftrightarrow T$ ↗

· RK.: if $\nabla\psi(x) = T_x$ is linear, then ψ is quadratic

· RK.: ψ convex $\Rightarrow$ ψ is a.e. differentiable $\Rightarrow$ $\nabla\psi$ exists α-a.e.

· Ex.: 1-D, $c = \|x-y\|^2$
$\alpha, \beta \in \mathcal{P}(\mathbb{R})$, α has density $\rho_\alpha > 0$. C_α the c.d.f., $Q_\alpha = C_\alpha^{-1}$ the quantile function.

· Prop.: $(C_\alpha)_\# \alpha = \mathcal{U}_{[0,1]}$, $(Q_\alpha)_\# \mathcal{U}_{[0,1]} = \alpha$.
show it

· Theo.: By Brenier, $Q_\beta \circ C_\alpha = T$ is the unique OT from α to β.

For ex., take $\alpha = \mathcal{N}(m_\alpha, \sigma_\alpha^2)$, $\beta = \mathcal{N}(m_\beta, \sigma_\beta^2)$
· Prop.: $T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$, $T_\# \alpha = \beta$
show it
· Cor.: T is the unique OT $W_2(\alpha,\beta) = \left((m_\alpha - m_\beta)^2 + (\sigma_\alpha - \sigma_\beta)^2\right)^{\frac{1}{2}}$

· RK.: $W_p(\alpha,\beta) := \left(\inf_{T_\#\alpha = \beta} \int \|x - Tx\|^p d\alpha(x)\right)^{1/p}$ (The Wasserstein (?) distance).

A note: Brenier
We can transform this into a PDE problem, but this isn't the right formulation for the problem we want to study.

· Ex.: Gaussians

$\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$, $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$, $\Sigma = U \operatorname{diag}(\sigma^2) U^T$

Ansatz: $T(x) = \hat{A}(x - m_\alpha) + m_\beta$ ($\hat{A}$?)

Lemma: $T_\# \alpha = \beta \Leftrightarrow A \Sigma_\alpha A^T = \Sigma_\beta$
Proof: $T_\# \alpha = \mathcal{N}(m, \Sigma)$, $m = \mathbb{E}[T(X)]$, $X \sim \alpha$, $TX \sim T_\# \alpha$

$m = m_\beta$, $\Sigma = \mathbb{E}[T(X)T(X)^T] = \dots = \mathbb{E}[(AX)(AX)^T] = A\,\mathbb{E}[XX^T]\,A^T = A \Sigma_\alpha A^T$

Lemma: $T(x) = A(x - m_\alpha) + m_\beta = \nabla\varphi(x)$, φ cvx $\Leftrightarrow$ $A^T = A$ and $A \succeq 0$

Cor: T is the OT $\Leftrightarrow$ $\begin{cases} A \Sigma_\alpha A = \Sigma_\beta \\ A \succeq 0 \end{cases}$ (Riccati equation)

Rk: If $A \geq 0$, $\exists!\ \sqrt{A} \geq 0$ s.t. $\sqrt{A}^2 = A$

Proof: $A = U \operatorname{diag}(\sigma^2)\, U^T$, $\sqrt{A} = U \operatorname{diag}(\sigma)\, U^T$

Prop.: $A \Sigma_\alpha A^T = \Sigma_\beta$ has a unique sol., $Q(\Sigma_\alpha, \Sigma_\beta)$.

Proof: $\left(\sqrt{\Sigma_\alpha}\, A \sqrt{\Sigma_\alpha}\right)\left(\sqrt{\Sigma_\alpha}\, A^T \sqrt{\Sigma_\alpha}\right) = \sqrt{\Sigma_\alpha}\, \Sigma_\beta \sqrt{\Sigma_\alpha} \iff \sqrt{\Sigma_\alpha}\, A \sqrt{\Sigma_\alpha} = \sqrt{\sqrt{\Sigma_\alpha}\, \Sigma_\beta \sqrt{\Sigma_\alpha}}$

(we are using α has a density $\iff \det(\Sigma_\alpha) \neq 0$).

$\iff A = \sqrt{\Sigma_\alpha}^{-1} \sqrt{\sqrt{\Sigma_\alpha}\, \Sigma_\beta \sqrt{\Sigma_\alpha}}\, \sqrt{\Sigma_\alpha}^{-1} = Q(\Sigma_\alpha, \Sigma_\beta)$

Def: If $AB = BA$, $G(A, B) := Q(A^2, B) = \sqrt{AB}$

- Def: $W_2^2(\alpha, \beta) = \int \|x - T^*(x)\|^2 \, d\alpha(x)$, T^* Brenier map.

Prop.: $W_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|_2^2 + \mathcal{B}^2(\Sigma_\alpha, \Sigma_\beta)$ (see exercise sheet 1)

$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta) := \operatorname{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\sqrt{\sqrt{\Sigma_\alpha}\, \Sigma_\beta \sqrt{\Sigma_\alpha}}\right)$.

Prop.: if $\Sigma_\alpha \Sigma_\beta = \Sigma_\beta \Sigma_\alpha$, $\mathcal{B}^2(\Sigma_\alpha, \Sigma_\beta) = \|\sqrt{\Sigma_\alpha} - \sqrt{\Sigma_\beta}\|_2^2 = \operatorname{Hell}^2(G_{\alpha^2}, G_{\beta^2})$,

$\operatorname{Hell}(G_{\alpha^2}, G_{\beta^2}) = \|G_\alpha - G_\beta\|_2$

Kantorovich Formulation

- Def: $P \in U(a, b)$ (set of transport plans) is a transport plan when $P \in \mathbb{R}_+^{n \times m}$ (positivity) and $\begin{cases} \forall i, \sum_j P_{ij} = a_j \\ \forall j, \sum_i P_{ij} = b_j \end{cases}$

(transporting mass)

$$U(a, b) := \left\{ P \in \mathbb{R}_+^{n \times m} : P \mathbb{1}_m = a,\ P^T \mathbb{1}_n = b \right\}$$

C_{ij} is the cost for a unit amount of mass to be transported $x_i \to y_j$

$$(K) \quad \min_P \left\{ \sum_{i,j} P_{ij} \cdot C_{ij} = \langle C, P \rangle : P \in U(a, b) \right\}$$

- Lemma: $U(a, b)$ is a non empty compact convex set

- Cor.: (K) always admits a solution. (continuous function on a compact set)

This is the origin of linear programming $\nearrow$ Kantorovich (theory, duality) $\searrow$ Dantzig (simplex algorithm)

We study a particular case: $n = m$, $a = b = \mathbb{1}$

$$(M) \quad \min_{\sigma \in \Sigma_n} \sum_i C(\sigma(i)), \quad \mathcal{P}_n = \left\{ P \in \mathbb{R}_+^{n \times n},\ \forall i,j,\ \begin{matrix} P[i,:] \\ P[:,j] \end{matrix} \text{ has a single non-zero el.} = 1 \right\} \quad \text{(permutation matrices)}$$

$$(M) = \min_{P \in \mathcal{P}_n} \langle P, C \rangle$$

$$\mathcal{B}_n := \left\{ P \in \mathbb{R}_+^{n \times n},\ P \mathbb{1}_n = P^T \mathbb{1}_n = \mathbb{1}_n \right\} \to (K) \quad \min_{P \in \mathcal{B}_n} \langle P, C \rangle$$

· **Theo.:** Birkhoff - Von Neumann

$(M) = (K)$, i.e. $\exists\ P$ solving (K) that is also a solution of (M).

· **Def.:** $P \in Extm(\mathcal{E}) \iff \left[P = \frac{Q+R}{2},\ Q, R \in \mathcal{E} \Rightarrow Q = R = P \right]$ (vertices)

· **Theo.:** If $\mathcal{E}$ is compact, then $Extm(\mathcal{E}) \neq \phi$.

· **Theo.:** (Krein - Milman)

$\mathcal{E}$ compact and convex $\Rightarrow \mathcal{E} = ConvHull(Extm(\mathcal{E}))$

· **Cor.:** $\left\{ \min_{P \in \mathcal{E}} \langle c, P \rangle \right\} \cap Extm(\mathcal{E}) \neq \phi$

Proof: $S := \operatorname{argmin} \{ \langle c, P \rangle,\ P \in \mathcal{E} \}$

S compact and convex. $S \supset Extm(S) \in Extm(\mathcal{E})$

$P \in Extm(S) \Rightarrow P \in Extm(\mathcal{E})$ $\quad P = \frac{Q+R}{2} \Rightarrow Q, R \in S$

· **Theo.:** (B-VN)

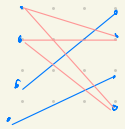
$\mathcal{P}_n = Extm(\mathcal{B}_n)$

Proof: i) $\mathcal{P}_n \subset Extm(\mathcal{B}_n)$ $\quad P_{ij} = \frac{Q+R}{2}_{ij},\ Q, R \in [0,1]^{n \times n} \rightarrow Q_{ij}, R_{ij} \in \{0,1\}$, because $P_{ij} \in \{0,1\}$

ii) $\mathcal{B}_n \setminus \mathcal{P}_n \overset{??}{\subset} \mathcal{B}_n \setminus Extm(\mathcal{B}_n)$

$P \in \mathcal{B}_n$, $P \notin \mathcal{P}_n$. Need to show $\exists\ Q \neq R$ $\quad P = \frac{Q+R}{2}$

By graph theory arguments, there is at least one cycle in the transport graph. $\{i_1, j_1, \dots, i_k, j_k, i_1\}$



$\theta \in \mathbb{E} = \min_{ij} \{ P_{ij},\ 1 - P_{ij} \}$

$Q_{ij} = \begin{cases} P_{ij} + \varepsilon & \text{if } (ij) \in \mathcal{A} \\ P_{ij} - \varepsilon & \text{if } (ij) \in \mathcal{B} \\ P_{ij} & \text{otherwise} \end{cases}$

$\mathcal{A} = \{ (i_s, j_s) \}$

$\mathcal{B} = \{ (j_s, i_{s+1}) \}$

Lecture 3

Kantorovich: $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$, $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$ $\rightarrow \min_{P \in \mathbb{R}_+^{n \times m}} \{ \langle C, P \rangle : P \mathbb{1}_m = a, \ P^T \mathbb{1}_n = b \}$

We will now look at the generalized problem.

· **Def.:** $\alpha \in \mathcal{P}(X)$, $\beta \in \mathcal{P}(Y)$. A coupling is $\pi \in \mathcal{P}(X \times Y)$ having marginals α and β, i.e. $\forall f \in \mathcal{C}(X)$, $\iint f(x) \, d\pi(x,y) = \int f(x) \, d\alpha(x)$, $\forall g \in \mathcal{C}(Y)$, $\iint g(y) \, d\pi(x,y) = \int g(y) \, d\beta(y)$

· **Def.:** $P_1 : (x,y) \in X \times Y \mapsto x \in X$, $P_2 : (x,y) \in X \times Y \mapsto y \in Y$
$\pi_1 := (P_1)_\# \pi$, $\pi_2 := (P_2)_\# \pi$

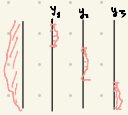
· **Prop.:** Coupling $\iff \pi_1 = \alpha$, $\pi_2 = \beta$
Do it

· **Prop.:** $\alpha \otimes \beta$ works

· **Prop.:** If marginals are discrete, $\begin{cases} \pi_2 = \beta \\ \pi_1 = \alpha \end{cases} \iff \pi = \sum_{ij} P_{ij} \delta_{(x_i, y_j)}$, $P \mathbb{1} = a$, $P^T \mathbb{1} = b$

$X = \bigsqcup_{j=1}^{n} \mathcal{X}_j$ disjoint, $T : x \in \mathcal{X}_j \rightarrow y_j$ piecewise constant.

· **Prop.:** iff. $\alpha(\mathcal{X}_j) = b_j$, $T_\# \alpha = \beta$



(Discrete-continuous)

· **Prop.:** Suppose $T_\# \alpha = \beta$, then $\pi = \overbrace{(Id, T)}^{\varphi}_\# \alpha$ is a valid coupling. $\varphi : x \in X \mapsto (x, T(x)) \in X \times Y$

(K) $\inf_{\pi \in \mathcal{P}(X \times Y)} \{ \iint c(x,y) \, d\pi(x,y) : \pi_1 = \alpha, \pi_2 = \beta \}$

· **Rk.:** if (*) (discrete case) holds, $\pi^* = \sum_{ij} P_{ij}^* \delta_{(x_i, y_j)}$ and P^* is a solution to the discrete Kantorovich problem.

· **Prop.:** If α and β have compact support (ex X, Y compact), then (K) has a solution.
Proof: When we introduce topology:
For the weak* topology $\mathcal{P}(X \times Y)$ is compact, $\pi \mapsto \iint c \, d\pi$ is continuous and $\alpha \otimes \beta$ is a valid coupling.

· **Prop.:** For a non compact space, $p \geq 1$, $X = Y$, $c(x,y) = d(x,y)^p$, $\int d(x_0, x)^p \, d\alpha(x) < +\infty$, $\int \dots d\beta < +\infty$, then (K) has a solution.

· **Theo.:** (Brenier)
$X = Y = \mathbb{R}^d$, $c(x,y) = \|x - y\|^2$, α has a density w.r.t. Lebesgue.
$\exists!$ solution $T = \nabla \varphi$ of Monge, $(\nabla \varphi)_\# \alpha = \beta$, $\pi = (Id, T)_\# \alpha$ is the unique sol. of (K).

· **Rk.:** $\pi = (Id, \pi)_\# \alpha \iff \iint c \, d\pi = \int c(x, T(x)) \, d\alpha(x)$

· **Rk.:** (K) $\leftrightarrow \inf_{(X,Y)} \{ \mathbb{E}[c(X,Y)], \ Law(X) = \alpha, \ Law(Y) = \beta \}$

· **Def.:** $X = Y$, $c(x,y) = d(x,y)^p$, $p \geq 1$.
Wasserstein distance: $W_p(\alpha,\beta) := \left(\inf_{\pi \in P(X\times X)} \iint d(x,y)^p \, d\pi(x,y) \right)^{1/p}$

· **Theo.:** $(P(X), W_p)$ is a complete separable metric space.

Proof: $W_p(\alpha,\beta) = 0 \Leftrightarrow \alpha = \beta$. $(\alpha,\beta,\gamma) \in \mathcal{M}(X)^3$, $W_p(\alpha,\gamma) \leq W_p(\alpha,\beta) + W_p(\beta,\gamma)$
We look into the discrete case.
$\alpha = \sum a_i \delta_{x_i}$, $\beta = \sum b_j \delta_{y_j}$, $\gamma = \sum c_k \delta_{z_k}$

$\sum d_{ij}^p P_{ij}^* = 0 \Leftrightarrow P = \mathrm{diag}(h)$. $P\mathbb{1} = a = h$, $P^T\mathbb{1} = b = h \Rightarrow a = b$

$a \xrightarrow{P} b \xrightarrow{Q} c$

Lemma (Gluing): $G_{ijk} = \dfrac{P_{ij} Q_{jk}}{b_j}$, $\sum_k G_{ijk} = P_{ij}$, $\sum_i G_{ijk} = Q_{jk}$ and $\sum_{jk} G_{ijk} = a_i$, $\sum_{ik} G_{ijk} = b_j$, $\sum_{ij} G_{ijk} = z_k$

Def. / Cor.: $R_{ik} = \sum_j G_{ijk}$ and we get $R\mathbb{1} = a$, $R^T\mathbb{1} = c$.
$\Leftrightarrow R = P \,\mathrm{diag}\left(\frac{1}{b}\right) Q$)

So: Take P, Q optimal $\rightarrow R$ coupling between a and c $(D_{ik} = d(x_i, z_k))$
$W_p(\alpha,\beta)^p \leq \left(\sum_{ik} D_{ik}^p \sum_j \frac{P_{ij} Q_{jk}}{b_j} \right)^{1/p} \leq \left(\sum_{ijk} (D_{ij} + D_{jk})^p \frac{P_{ij} Q_{jk}}{b_j} \right)^{1/p}$.
We use Minkowski $\left(\sum_j \lambda_j (a_j + b_j)^p\right)^{1/p} \leq \left(\sum_j \lambda_j a_j^p\right)^{1/p} + \left(\sum_j \lambda_j b_j^p\right)^{1/p}$
So $W_p(\alpha,\beta)^p \leq \underbrace{\left(\sum_{ij} D_{ij}^p P_{ij}\right)^{1/p}}_{= W_p(\alpha,\beta)} + \left(\sum_{jk} D_{jk}^p G_{ijk}\right)^{1/p} = W_p(\alpha,\beta) + W_p(\beta,\gamma)$.

· **Thm.:** $\pi \in P(X\times X)$, $\xi \in P(X\times X)$ $\Leftrightarrow \exists \varrho \in \mathcal{P}(X\times X\times X)$, $\varrho_{12} = (P_{12})_{\#}\varrho = \pi$, $\varrho_{23} = (P_{23})_{\#}\varrho = \xi$

· **Def.:** (Total variation)
$\|\alpha - \beta\|_{TV} = |\alpha - \beta|(X)$

· Interesting comparison:
$(x_k) \to x$. $\|\delta_{x_k} - \delta_x\|_{TV} = 2 \not\to 0$, $W_p(\delta_{x_k}, \delta_x) = d(x_k, x) \to 0$

· **Def.:** Convergence in TV is strong convergence.

· **Def.:** $\alpha_k \rightharpoonup^* \alpha \Leftrightarrow \forall f \in C_b(X)$, $\int f(x)\, d\alpha_k(x) \to \int f(x)\, d\alpha(x)$

· **Prop.:** $\|\alpha\|_{TV} = \sup_{\|f\|_\infty \leq 1} \int f(x)\, d\alpha(x)$

· **Theo.:** Wasserstein is cv in law.
If X compact, $\alpha_k \rightharpoonup^* \alpha \Leftrightarrow W_p(\alpha_k, \alpha) \xrightarrow{k\to\infty} 0$
In general $W_p(\alpha_k, \alpha) \to 0 \Leftrightarrow \begin{cases} \alpha_k \rightharpoonup^* \alpha \\ \int d(x_0,x)^p \, d\alpha_k \to \int d(x_0,x)^p \, d\alpha \end{cases}$

· **Theo.:** If $\mathbb{E}[\|X\|^3] < +\infty$, then $W_p(\alpha_n, \mathcal{N}(0,\Delta)) \leq \dfrac{C}{\sqrt{n}}$ (α^* the law of $\frac{\bar{X}_n - \mu}{\sigma^2}$)

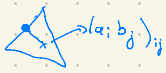
Lecture 4

We regularise the problem:

$$\min_{P \in \mathbb{R}^{n\times m}_+} \{ \langle C, P \rangle + \varepsilon H(P) : P\mathbb{1} = a, \; P^T \mathbb{1} = b \} \qquad \text{(Schrödinger problem)}$$

- **Def:** Shannon entropy: $H(P) := \sum_{ij} P_{ij} \log(P_{ij})$
- **Rk:** As ε increases our graph becomes more and more connected.
- **Prop:** P is the solution of (Schr) iff. $\begin{cases} P\mathbb{1} = a \\ P^T\mathbb{1} = b \\ \exists u \in \mathbb{R}^n_+, v \in \mathbb{R}^m_+, \; P = \mathrm{diag}(u) K \, \mathrm{diag}(v) \end{cases}$
- **Def:** $K = \exp(-\frac{C}{\varepsilon}) = (\exp(-\frac{c_{ij}}{\varepsilon}))_{ij}$
- **Rk:** For W_2, K is the Gaussian kernel.

Proof: Step 1: $P_{ij} > 0 \; \forall i,j$. Suppose $P_{i_0 j_0} = 0$.



$P^t_{ij} = (1-t) P_{ij} + t(a_i b_j)$, $g(t) = \langle C, P^t \rangle + \varepsilon H(P^t)$

Prop: $g'(0) = -\infty$. This implies that there is a better solution than the one with $P_{i_0 j_0} = 0$.

Now: $\min_{P \geq 0} \{ \langle C, P^t \rangle + \varepsilon H(P) : P\mathbb{1} = a, \; P^T\mathbb{1} = b \}$ $\begin{matrix} f \in \mathbb{R}^n \\ g \in \mathbb{R}^m \end{matrix}$

$$\mathcal{L}(P, f, g) = \langle C, P \rangle + \varepsilon H(P) + \underbrace{\langle a - P\mathbb{1}, f \rangle}_{\langle a, f \rangle - \langle P, f \mathbb{1}^T \rangle} + \underbrace{\langle b - P^T\mathbb{1}, g \rangle}_{\langle b, g \rangle - \langle P, \mathbb{1} g^T \rangle}$$

$$\frac{\partial \mathcal{L}}{\partial P} = 0 = C + \varepsilon(\log(P) + \mathbb{1}) - f\mathbb{1}^T - \mathbb{1}g^T = 0$$

$\Leftrightarrow P_{ij} = \exp(-\frac{c_{ij}}{\varepsilon}) \cdot \exp(f_i/\varepsilon) \cdot \exp(g_j/\varepsilon) \cdot e$ and get the formula.

$\odot$ is element-wise multiplication

$$\begin{cases} (R) \; u \odot (Kv) = a \Leftrightarrow u = \frac{a}{Kv} \\ (C) \; v \odot (K^T u) = b \Leftrightarrow v = \frac{b}{K^T u} \end{cases} \qquad \text{(element-wise division)}$$

Many names for the algorithm: Sinkhorn, Iterative scaling, IPFP

$v_{t=0} = \mathbb{1} \to u_{t+1} = \frac{a}{K v_t} \to v_t = \frac{b}{K^T u_{t+1}}$

$O(T \cdot n^2)$, T = # of iterations

Convergence? (spoiler: yes)

• **Prop.**: to compute the Wasserstein distance to precision δ ($\varepsilon \sim \delta$ up to $\log(d)$) and even $\frac{1}{\varepsilon^2} \approx \frac{1}{\delta^2}$ iterations of Sinkhorn. ($|\langle C, P_\varepsilon \rangle - \langle C, P^* \rangle| \sim \delta$).

• **Def.**: $KL(P|Q) = \sum_{ij} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right)$. $H(P) = KL(P|\mathbb{1})$

• **Prop.**: $KL(P|Q) \geq 0$, $KL(P|Q) = 0 \Rightarrow P = Q$

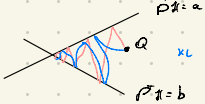
• **Prop.**: $P\mathbb{1} = a$, $P^T\mathbb{1} = b$, $KL(P|a\otimes b) = KL(P|a'\otimes b') - KL(a|a') - KL(b|b')$
Therefore, taking $a' = \mathbb{1}$, $b' = \mathbb{1}$, $KL(P|a\otimes b) = H(P) + cst$

$(Sch) \Leftrightarrow \min_{P \geq 0} \left(\langle C, P \rangle + \varepsilon KL(P|a\otimes b) : P\mathbb{1} = a, \quad P^T\mathbb{1} = b \right)$

• **Prop.**: $P \xrightarrow{\varepsilon \to 0} a \otimes b$

• **Prop.**: $\langle C, P \rangle + \varepsilon KL(P|a\otimes b) = \varepsilon\, KL\left(P \,|\, e^{-C/\varepsilon} \odot (a\otimes b)\right) + cst$

This is all a story of projection on convex sets:



$\text{Proj}^{KL}_{\mathcal{C}}(P) = \inf_{P' \in \mathcal{C}} KL(P|P')$, $P_0 = Q$, $P_{t+\frac{1}{2}} = \text{Proj}^{KL}_{\mathcal{C}_1}(P_t)$, $P_{t+1} = \text{Proj}^{KL}_{\mathcal{C}_2}(P_{t+\frac{1}{2}})$

• **Theo.**: (Bregman)
$P_t \to \mathcal{C}_1 \cap \mathcal{C}_2$ (the constraints). If $\mathcal{C}_1$ and $\mathcal{C}_2$ are affine, $P_t \to \text{Proj}_{\mathcal{C}_1 \cap \mathcal{C}_2}(Q)$

• **Prop.**: This is Sinkhorn.

φ strictly convex,
$B_\varphi(x|y) = \varphi(x) - \varphi(y) - \langle y - x, \nabla\varphi(x) \rangle$
we can build many things from that: $\varphi(s) = \frac{\|x\|^2}{2} \Rightarrow B_\varphi(x|y) = \frac{1}{2}\|x-y\|^2$, $\varphi(s) = H(s) \Rightarrow B_\varphi = KL$

• **Def.**: $\mathcal{E} := \mathbb{R}^{+d}_* / \sim$, $u \sim v \Leftrightarrow \exists \lambda > 0 : u = \lambda v$
$d_H(u,u') = \left\| \log\left(\frac{u}{u'}\right) \right\|_V$, $\|z\|_V = \max(z) - \min(z)$

• **Theo.**: $(\mathcal{E}, d_H)$ is a complete metric space.

• **Theo.**: (Birkhoff contraction)
$K : \mathcal{E} \to \mathcal{E}$ linear. $\exists \eta < 1$, $d_H(Ku, Ku') \leq \eta\, d_H(u,u')$

• **Prop.**: $d_H\left(\frac{a}{u}\cdot\frac{a}{u'}\right) = d_H(u,u')$

• **Cor.**: $d_H(S(u), S(u')) \leq \eta^2 d_H(u,u')$ ($S(u)$ means one iteration of Sinkhorn)

• **Prop.**: $\eta \sim \exp\left(-\frac{2\|C\|_\infty}{\varepsilon}\right) \xrightarrow[\varepsilon \to 0]{} 1$

$\alpha, \beta \in \mathcal{M}^+(\mathcal{X}), \mathcal{M}^+(\mathcal{Y})$, $\inf_{\pi} \left\{ \iint_{\mathcal{X}\times\mathcal{Y}} c \, d\pi + \varepsilon KL(\pi \mid \alpha \otimes \beta) \mid \pi_1 = \alpha, \pi_2 = \beta \right\}$

• Theo: If c continuous and $\mathcal{X}, \mathcal{Y}$ compact, then (Schrödinger) has a unique solution π_ε.

$$\left(\frac{d\pi}{d\alpha d\beta}(x, y) = u(x) \exp\left(-\frac{c(x,y)}{\varepsilon}\right) v(y) \right.$$

• Def: $X, Y \sim \pi$, $I(X, Y) := KL(\pi \mid \pi_1 \otimes \pi_2)$

• Prop: $I(X, Y) = 0$ iff. $X \perp\!\!\!\perp Y$

(Sch) $\inf_{(X,Y)} \left\{ E[c(X, Y)] + \varepsilon I(X, Y) : X \sim \alpha, Y \sim \beta \right\}$

Lecture 5

Kernels: $X = Y$, $K(x,y) \in \mathbb{R}$ that is positive definite.

- **Def.:** $K(x,y)$ is universal iff: $\mathrm{span}\{x \mapsto K(x,y) : y \in X\}$ is dense in $\mathcal{C}(X_0, \mathbb{R})$ $\forall X_0 \in X$ compact, for uniform convergence on $\mathcal{C}$.

- **Ex.:** $K(x,y) = \exp\left(-\frac{\|x-y\|^p}{\sigma^p}\right)$

- **RK:** PD Kernel $\Leftrightarrow$ $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$

- **Def:** Maximum mean discrepancy (MMD)
$\|\alpha - \beta\|_K^2 = \iint K(x,y)\, d(\alpha-\beta)(x)\, d(\alpha-\beta)(y)$

- **RK:** $K(x,y) = h(x-y) \rightarrow \int h(x-y)\, d\gamma(y) = (h * \gamma)(x)$

$\iint K\, d\gamma \otimes \gamma = \langle h * \gamma, \gamma \rangle$

- **Theo.:** (Def. 4)
If K is conditionally p.d. and universal, then $\|\cdot\|_K$ is a norm on $\mathcal{P}(X)$.

$\alpha_s \rightharpoonup \alpha \Leftrightarrow \|\alpha - \alpha_s\|_K \rightarrow 0$

- **Discrete case:**

$\alpha = \frac{1}{n}\sum \delta_{x_i}$, $\beta = \frac{1}{m}\sum \delta_{y_j}$, $\|\alpha - \beta\|_K^2 = \|\alpha\|_K^2 + \|\beta\|_K^2 - 2\langle \alpha, \beta \rangle_K = \frac{1}{n^2}\sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{m^2}\sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{nm}\sum_{i,j} k(x_i, y_j)$

- $W_c^\varepsilon(\alpha,\beta) = \min_{\substack{\pi_1 = \alpha \\ \pi_2 = \beta}} \left(\iint c\, d\pi + \varepsilon\, KL(\pi \,|\, \alpha \otimes \beta) \right) \xrightarrow{\varepsilon \to 0} W_c(\alpha,\beta)$

⚠ $W_c^\varepsilon(\alpha,\alpha) > 0$.

$\pi_\varepsilon = \mathrm{argmin}\, W_c^\varepsilon(\alpha,\beta)$

- **Prop:** $\pi_\varepsilon \xrightarrow{\varepsilon \to 0} \alpha \otimes \beta$, so $W_c^\varepsilon(\alpha,\beta) \xrightarrow{\varepsilon} \iint c(x,y)\, d\alpha \otimes \beta(x,y)$

- **Def.:** $K_c(x,y) = -c(x,y)$

- **Def:** Sinkhorn divergence (Thibault Séjourné)

$\bar{W}_c^\varepsilon(\alpha,\beta) = W_c^\varepsilon(\alpha,\beta) - \frac{1}{2} W_c^\varepsilon(\alpha,\alpha) - \frac{1}{2} W_c^\varepsilon(\beta,\beta)$

- **Prop:** $\bar{W}_c^\varepsilon(\alpha,\beta) \xrightarrow{\varepsilon \to 0} W_c(\alpha,\beta) \geq 0$

$\bar{W}_c^\varepsilon(\alpha,\beta) \xrightarrow{\varepsilon \to \infty} -\langle \alpha, \beta \rangle_{K_c} + \frac{1}{2}\|\alpha\|_{K_c}^2 + \frac{1}{2}\|\beta\|_{K_c}^2 = \frac{1}{2}\|\alpha - \beta\|_{K_c}^2$

• **Theo.:** If $\exp(-\frac{c}{\varepsilon})$ is a valid kernel, then $\bar{W}_c^\varepsilon \geq 0$, $\bar{W}_c^\varepsilon(\alpha,\beta) = 0 \Leftrightarrow \alpha = \beta$,
$\alpha_s \xrightarrow{c} \alpha \Leftrightarrow \bar{W}_c^\varepsilon(\alpha_s, \alpha) \to 0$.

• **Theo.:** $d = W_p$, $\Delta_n \sim \frac{1}{n^{1/d}}$, $\Delta_n = \mathbb{E}\left[|d(\alpha,\beta) - d(\hat{\alpha},\hat{\beta})|\right]$

where $\hat{\alpha}, \hat{\beta}$ are random measures sampling from α, β.

If $d = MMD$, $\Delta_n \leq \frac{C_d}{\sqrt{n}}$

If $d = W_c^\varepsilon$, $\Delta_n \leq C_d\left(1 + \varepsilon^{-\frac{d}{2}}\right) \cdot \frac{1}{\sqrt{n}}$

• **Theo.:** if $c(x,y)$ is lips., X, Y compact, then $\min_{\substack{\pi_1 = \alpha \\ \pi_2 = \beta}} \int\int c\, d\pi = \sup_{\substack{f \in \mathcal{C}(X) \\ g \in \mathcal{C}(Y)}} \left\{ \int f(x)\, d\alpha(x) + \int g(y)\, d\beta(y) \;:\; \forall x,y,\ f(x) + g(y) \leq c(x,y) \right\}$

• **Def.:** $g \in \mathcal{C}(Y) \longrightarrow g^c(x) = \min_{y \in Y} \left(c(x,y) - g(y)\right) \in \mathcal{C}(X)$ (from $\mathcal{C}(X)$ to $\mathcal{C}(Y)$)

so $\sup_{f \in \mathcal{C}(Y)} \left(\int f\, d\beta \;:\; f(x) + g(y) \leq c(x,y) \right) = \sup_g \left\{ \int g(y)\, d\beta(y) \;:\; g(y) \leq f^{\bar{c}}(y) \right\}$

• **Prop.:** $g = f^{\bar{c}}$ β-a.e.

Summary: f fixed $\to$ $f^{\bar{c}}$ sol. $\quad$ g fixed $\to$ g^c solution.

• **Def.:** Legendre transform

$$f^*(y) = \sup \langle x, y \rangle - f(x)$$

• **Def.:** c-concave functions $:= \{ f : \exists g,\ f = g^c \}$ $\quad$ ($c(x,y) = \langle x, y \rangle \to$ concave $=$ c-concave)

• **Prop.:** f c-concave $\Rightarrow f^{\bar{c}\,c} = f$

Lecture 6

1 mistakes because of an accident.

- **Def.:** $\delta f(\alpha) : \mathcal{M}^{\mathcal{U}} \to \mathbb{R}$

$$f(\alpha + \varepsilon \xi) = f(\alpha) + \varepsilon \underbrace{\langle \xi, \delta f(\alpha) \rangle}_{\int \delta f(\alpha)(x)\, d\xi(x)} + o(\varepsilon)$$

$$\frac{\partial \alpha}{\partial t} = -\delta f(\alpha_t)$$

Ex.: $f(\alpha) = \int h(x)\, d\alpha(x)$, $\delta f(x) = h \Rightarrow \frac{\partial \alpha}{\partial t} = -h$

$$f(\alpha) = \int |\nabla \alpha(x)|^2 dx = -\langle \Delta \alpha, \alpha \rangle \Rightarrow \delta f(\alpha) = -\Delta \alpha(x)$$

Fact: (Classical flow of Dirichlet is the heat equation $\frac{\partial \alpha}{\partial t} = \Delta \alpha$.

$$\alpha_{t+\tau} = \min_{\alpha} \frac{1}{2\tau} W_2^2(\alpha_t, \alpha) + f(\alpha)$$

$\tau \to 0$ gives the Wasserstein gradient flow of f.

$\alpha = \delta_x$, $F(x) = f(\delta_x)$

$$x(t+\tau) = \min_{x} \frac{1}{2\tau} \|x - x_t\|^2 + \tau F(x) \quad \text{(Implicit Euler)}$$

The opt. cond. is $(x_{t+\tau} - x_t) + \tau \nabla F(x_{t+\tau}) = 0$

$$(IE) \quad x_{t+\tau} = (Id + \tau \nabla F)^{-1}(x_t) \quad \| \quad \underset{\text{GD}}{(EE)} \quad x_{t+\tau} = (Id - \tau \nabla F)(x_t)$$

- **Theo.:**

Under appropriate hypothesis, the W_2-GF (Gradient Flow) is the solution of:

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}\left(\alpha_t \left(-\nabla_W f(\alpha_t)\right)\right) = 0$$

where $\nabla_W f(\alpha) = \nabla_{\mathbb{R}^d}[\delta f(\alpha_t)]$

Linear f: $f(\alpha) = \int h(x)\, d\alpha(x)$, $\delta f(\alpha) = h$, $\nabla_W f(\alpha)(x) = \nabla h(x)$

$$\frac{\partial \alpha}{\partial t} - \operatorname{div}(\nabla h\, \alpha) = 0 \to \dot{x} = -\nabla h(x)$$

Shannon: $f(\alpha) = \int \log\left(\frac{d\alpha}{dx}(x)\right) d\alpha(x)$, $\delta f(\alpha) = \log(\alpha(x)) + 1$, $\nabla_{\mathbb{R}^d} \delta f(\alpha) = \frac{\nabla \alpha}{\alpha}(x)$

$$\boxed{\frac{\partial \alpha}{\partial t} = \Delta \alpha} \quad \text{(heat equation)}$$

$\dots \rightsquigarrow \inf_{\nu} \left\{ f(\alpha_t) + \tau \int \left[\frac{|v_x|^2}{\tau} + \langle v_x, \nabla \frac{\delta f}{\delta \alpha}(\alpha_t) \rangle \right] d\alpha_t + o(\tau) \right\}$

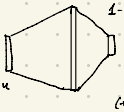
More: $f(\alpha) = \int g\left(\frac{d\alpha}{dx}\right) dx$, $\frac{\delta f}{\delta \alpha} = g'\left(\frac{d\alpha}{dx}\right)$, $\nabla_W f(\alpha) = g'(|\alpha'(x)|) \cdot \nabla \alpha(x)$

$\frac{\partial \alpha}{\partial t} + \operatorname{div}\left(\nabla \alpha(x) \cdot h(\alpha) \right) = 0$, where $h(s) = g'(s) \cdot s$

- **Def.:** f is geodesically convex when $\forall \alpha, \beta$, $t \mapsto f\left((1-t)\alpha + t\beta\right)$ is convex

- **Theo.:** $\int h \, d\alpha$ geod. convex $\Leftrightarrow$ h convex
  $\rightarrow \int \alpha \log \alpha$ is geod. convex

- **Theo.:** (McCann)
  $f(\alpha) = \int g\left(\frac{d\alpha}{dx}\right)$ is geod. convex if $g(0) = 0$, g convex ↗, $g(s^{-d}) s^d$ is ↘ in s.



1-hidden layer

$f_\theta(u) = \frac{1}{n} \sum_{i=1}^{n} a_i \sigma(\langle u, \omega_i \rangle) = \frac{1}{n} A \sigma(W u)$, $\theta_i = (\omega_i, a_i)$, $\alpha = \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i}$

(*) $f_\alpha(u) = \int \varphi(\theta, u) \, d\alpha(\theta)$, $\varphi(\theta, x) = a \, \sigma(\langle \omega, u \rangle)$

- **Prop.:** $E(\alpha) = \text{cst} + \int h(\theta) \, d\alpha(\theta) + \iint K(\theta, \theta') \, d\alpha(\theta) \, d\alpha(\theta')$, $h(\theta) = \frac{2}{N} \sum_k y_k - \varphi(\theta, u_k)$
  $K(\theta, \theta') = \frac{1}{N} \sum_k \varphi(\theta, u_k) \varphi(\theta', u_k)$ (we look for $\min_\alpha E(\alpha)$)

- **Theo.:**
  If $\alpha_{t=0}$ has enough particles (n large) and $\alpha_t \to \alpha_\infty$, then α_∞ is a global min.